Patients prescription churn case study

Gabriel Scozzarro

26/4/2021

1.0 Introduction

This case study origin from the problem that many retail shops has, that is customer retention. One way to tackle this problem is understand when an already acquired customer is churning.

But what is customer churn? Customer churn refers to when a customer (player, subscriber, user, etc. depending on industry) ceases his or her relationship with the company. Online businesses typically treat a customer as churned once a particular amount of time has elapsed since the customer's last interaction with the site or service. The full cost of customer churn includes both lost revenue and the marketing costs involved with replacing those customers with new ones. Reducing customer churn is a key business goal of every online business.

The ability to predict that a particular customer is at a high risk of churning, while there is still time to do something about it, represents a huge additional potential revenue source for every online business. Besides the direct loss of revenue that results from a customer abandoning the business, the costs of initially acquiring that customer may not have already been covered by the customer's spending to date. (In other words, acquiring that customer may have actually been a losing investment.) Furthermore, it is always more difficult and expensive to acquire a new customer than it is to retain current paying customers.

1.1 Scope of work & Business task

In this case study I will focus on predicting customer churn for a pharmacy retail group with many shops around the country. Furthermore a complete data analysis and exploration is performed, including insight on the prediction power of each data features which can be exploited to design customer retantion strategies.

2.0 Data

The data set used is avaible on **Kaggle** under the name of **Customer Churn Prediction || Pharmaceutical Data**. This data set is provided in csv format and has a Kaggle usability score of 8.8 which mean it was already cleaned and feature was drescribed for easy understanding. The data set was created in Feb 2021 and it contain pharmacies transaction and bill for a period of 5 month from the end of Apr 2019 to Sep 2019. In addition to the main data set it also provided a test dataset to test the prediction efficacy.

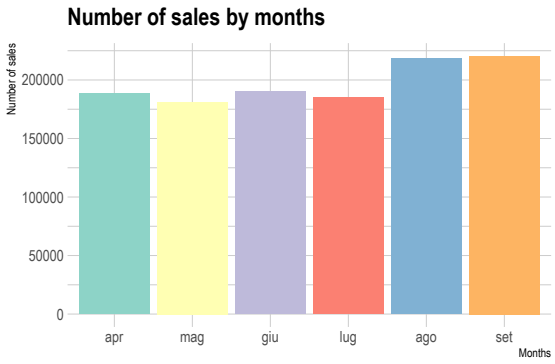
3.0 Tools and process

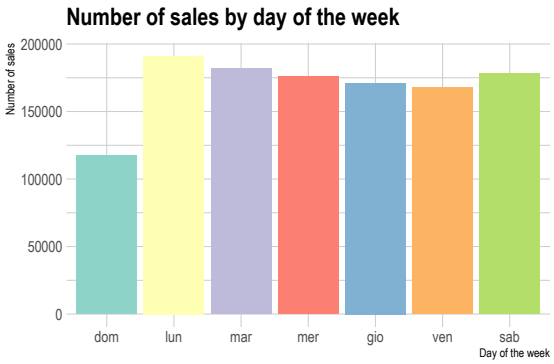
The analysis was performed using R coding language. A complete list of R packages, a data log and the source code are all avaible on the project github repository at this **link**.

4.0 Analysis

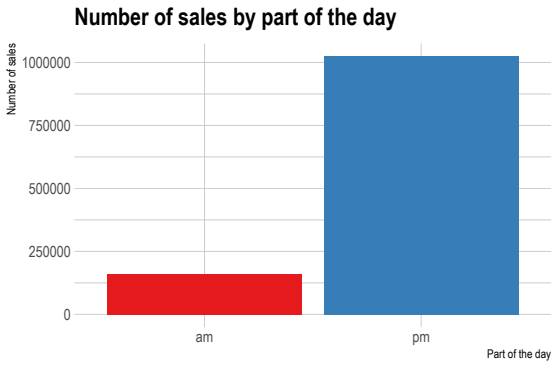
Data contains information about 1184025 transactions made in 43 stores, by 0 distinct users. The average receipt value was 306.7370216$.

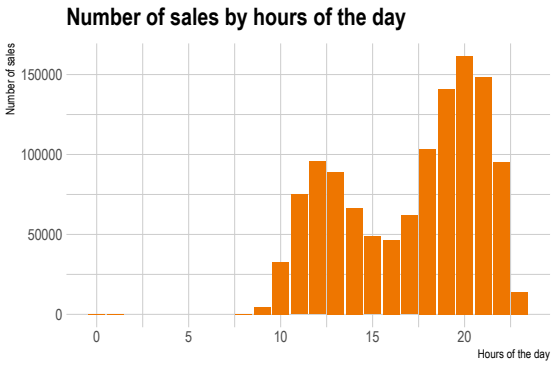
Tha sales analysis showed a steady number of sales from april to july with a costant increase in august and september. The highest number of sales was made on monday, while from tuesday to saturday we observed a slightly lower and constant value. Sunday has the worst sales rate, almost halved compared to the others.



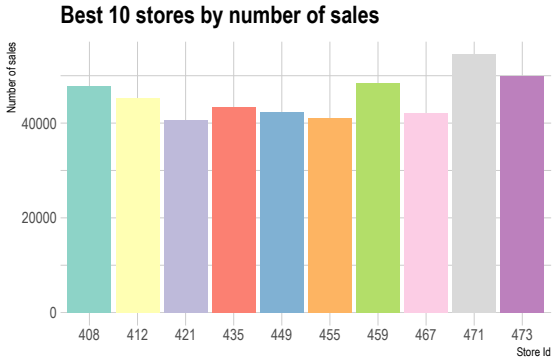


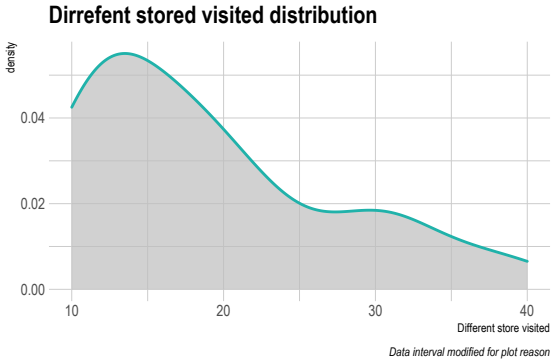
The favorite part of the day was the post meridian (pm) with a concentration on the between 17:00 and 22:00.



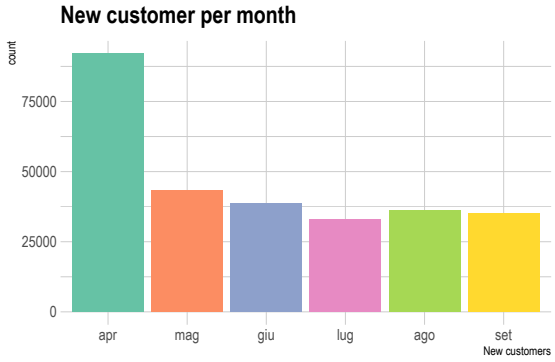


In the following plot are showed the top 10 store by number of sales.

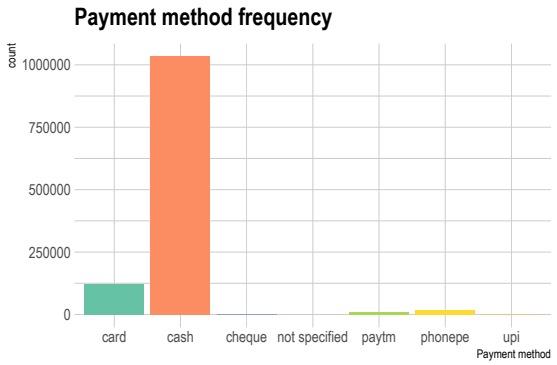


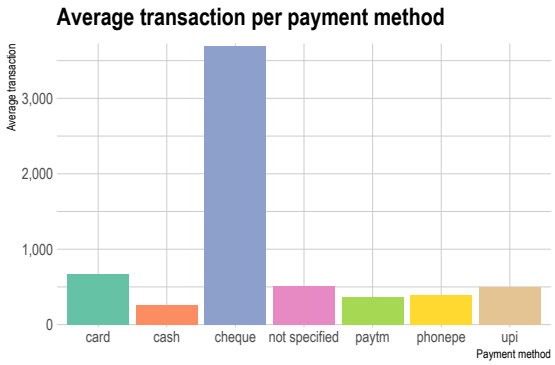


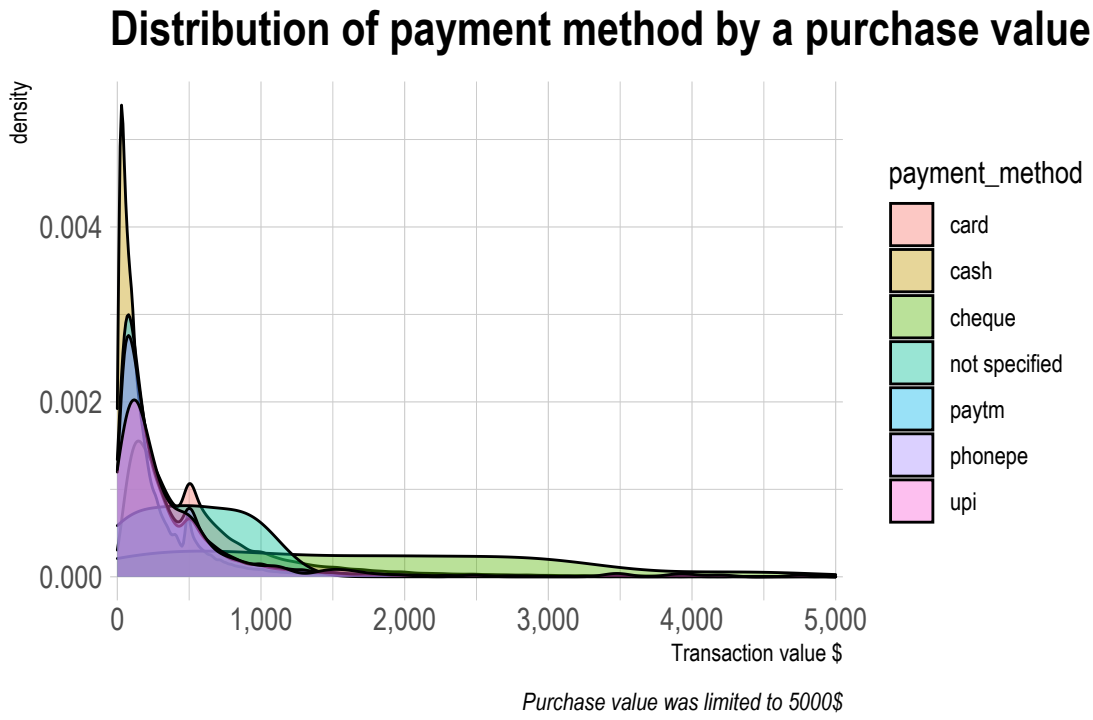
We can observe that the majority of the customer did the first purchase in april, this mean that the customer was acquired in april or previously. Not taking in consideration april the best month for acquisition was may the other has a steady increase. Overall the undisputed favorite payment method was cash. We note also that many customer made purchases in different store denoting a fidelization of that customer.



Overall the undisputed most used payment method was cash, but we can see that as the purchase value rise the payment method change.







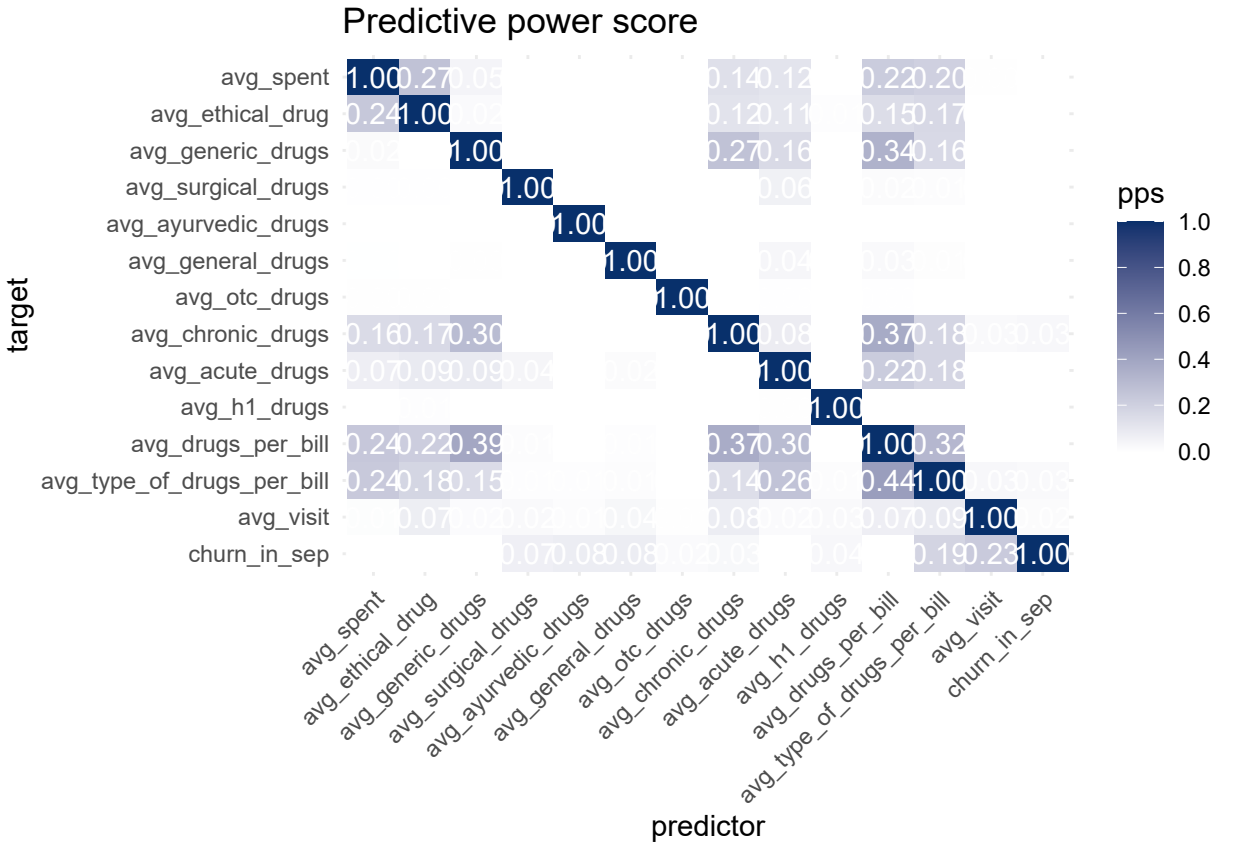
5.0 Prediction model

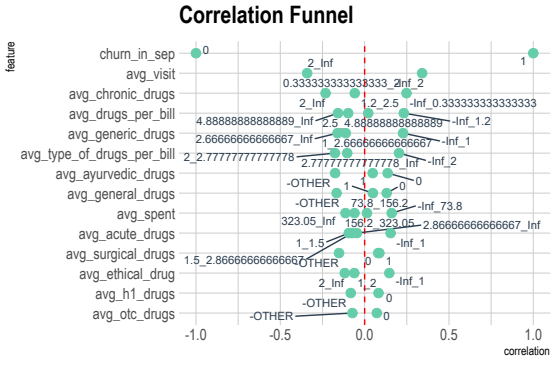
The desired prediction model need to predict the probability of each user to churn in september 2019. To build a prediction model based on this data set, a process called feature selection and engineering was necessary. The data frame created for train and validate the model was used also to evaluate the prediction power, correlation and importance of the features created. As target value a specific feature was created with the name of 'churn_in_sep' and can assume a logic value of 0 if the user stayed also in september and 1 it churn.

After this process, from the main data set a train and validation data frame were created. The evaluation of the model was made using AUC metrics.

Starting from the original data set, the target value inside the data frame prepared for the train and validation of the model has the following proportion: 30% of all users stays also in september, 70% of them churn. This mean that the original data was significantly unbalanced and a perfect prediction system is not possible. In the future work section I will present possible solution to this problem.

As shown in the following plot the most useful feature are the average quantity of visit made by the user and the the average quantity of drug for chronic diseases witch has a lot of sense since chronic diseases requires complex and prolonged treatment.





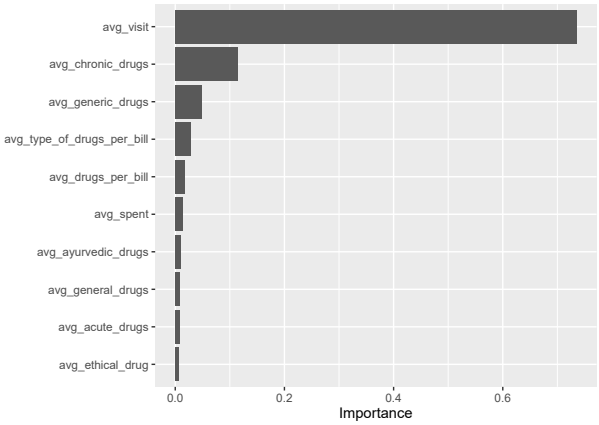


Table 1: Auto machine learning leaderboard

model_id	auc	logloss	aucpr	mean_per_class_error	rmse	mse
StackedEnsemble_AllModels_AutoML_20210504_152654	0.8192495	0.4621520	0.9028478	0.3132273	0.3867688	0.1495901
StackedEnsemble_BestOfFamily_AutoML_20210504_152654	0.8179639	0.4634521	0.9020051	0.3204532	0.3874177	0.1500925
GBM_5_AutoML_20210504_152654	0.8177569	0.4636521	0.9017971	0.3178905	0.3874712	0.1501339
GBM_4_AutoML_20210504_152654	0.8168636	0.4649724	0.9015561	0.3135111	0.3881203	0.1506374
GBM_3_AutoML_20210504_152654	0.8166248	0.4653212	0.9014328	0.3221382	0.3883083	0.1507833

For the prediction model, the approach taken was a generalist one, more a starting point rather than a complex and definitive solution. Using the **H2o.ai** platform in auto machine learning configuration, I was able to create different models using a variety of machine learning algorithms.

As expected from the unbalanced data frame the prediction model is more sensible in predict customer that will churn. To evaluate the best model, the metric AUC was used. AUC stands for Area under the ROC Curve (showed in the picture below) and it measures the entire two-dimensional area underneath the entire ROC curve (think integral calculus). AUC provides an aggregate measure of performance across all possible classification thresholds. One way of interpreting AUC is as the probability that the model ranks a random positive example more highly than a random negative example.

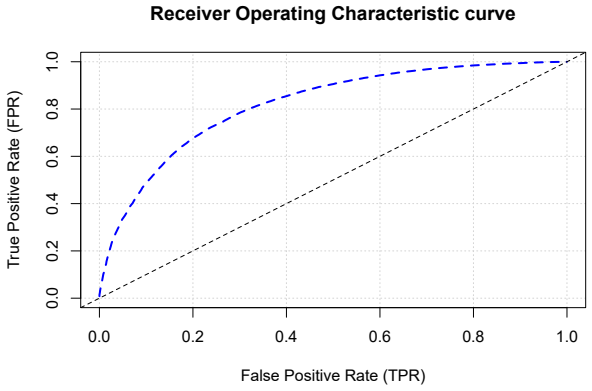
According to this metric the best model created was an stacked ensemble generated from overlap different algorithms.

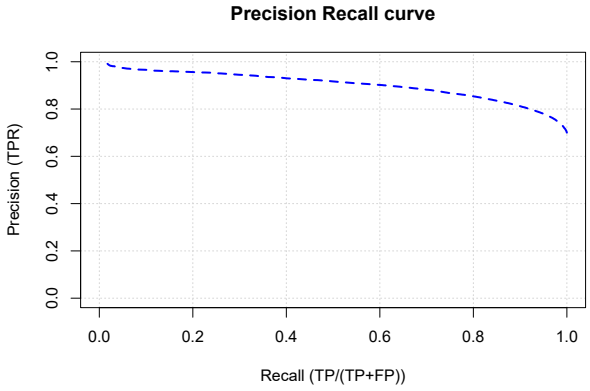
Table 2: Best model maximum metrics

metric	threshold	value	idx
max f1	0.4094247	8.570049e-01	285
max f2	0.1853513	9.244954e-01	363
max f0point5	0.6736565	8.430278e-01	178
max accuracy	0.4790741	7.845065e-01	258
max precision	0.9761995	9.970458e-01	2
max recall	0.0404970	1.000000e+00	399
max specificity	0.9806903	9.999086e-01	0
max absolute_mcc	0.6283476	4.655104e-01	197
max min_per_class_accuracy	0.7164893	7.405139e-01	159
max mean_per_class_accuracy	0.7371287	7.421740e-01	150
max tns	0.9806903	2.187200e+04	0
max fns	0.9806903	5.062600e+04	0
max fps	0.0404970	2.187400e+04	399
max tps	0.0404970	5.104700e+04	399
max tnr	0.9806903	9.999086e-01	0
max fnr	0.9806903	9.917527e-01	0
max fpr	0.0404970	1.000000e+00	399
max tpr	0.0404970	1.000000e+00	399

Table 3: Best model confusion matrix

	0	1	Error	Rate
0	1498	1551	0.5086914	=1551/3049
1	405	6514	0.0585345	=405/6919
Totals	1903	8065	0.1962279	=1956/9968





6.0 Conclusion

Churn Prediction was done performed and results were acceptable. The data set creation and feature engineering were the complicated part of this project. Higher accuracy cannot be achieved simply in this type of data set. Training the previous month data set, we can predict the next months customers status in churn prediction.

7.0 Future work

Since the unbalanced nature of the data provide and the impossibility to gather new, a synthetic data augmentation process can be applied using for example ROSE technique.